**Fikret Inal**

Department of Chemical Engineering,
Izmir Institute of Technology,
Gulbahce-Urla, Izmir, Turkey

Research Article

# Artificial Neural Network Prediction of Tropospheric Ozone Concentrations in Istanbul, Turkey

Tropospheric (ground-level) ozone has adverse effects on human health and environment. In this study, next day's maximum 1-h average ozone concentrations in Istanbul were predicted using multi-layer perceptron (MLP) type artificial neural networks (ANNs). Nine meteorological parameters and nine air pollutant concentrations were utilized as inputs. The total 578 datasets were divided into three groups: training, cross-validation, and testing. When all the 18 inputs were used, the best performance was obtained with a network containing one hidden layer with 24 neurons. The transfer function was hyperbolic tangent. The correlation coefficient ($R$), mean absolute error (MAE), root mean squared error (RMSE), and index of agreement or Willmott's Index ($d_2$) for the testing data were 0.90, 8.78 $\mu g/m^3$, 11.15 $\mu g/m^3$, and 0.95, respectively. Sensitivity analysis has indicated that the persistence information (current day's maximum and average ozone concentrations), NO concentration, average temperature, $PM_{10}$, maximum temperature, sunshine time, wind direction, and solar radiation were the most important input parameters. The values of $R$, MAE, RMSE, and $d_2$ did not change considerably for the MLP model using only these nine inputs. The performances of the MLP models were compared with those of regression models (i.e., multiple linear regression and multiple non-linear regression). It has been found that there was no significant difference between the ANN and regression modeling techniques for the forecasting of ozone concentrations in Istanbul.

**Keywords:** Artificial neural network (ANN); Istanbul; Multi-layer perceptron (MLP); Ozone; Regression

*Received:* April 13, 2010; *revised:* June 29, 2010; *accepted:* July 30, 2010

**DOI:** 10.1002/clen.201000138

## 1 Introduction

Ozone ($O_3$) is a secondary pollutant and formed in the lower atmosphere (troposphere) by the complex reactions of nitrogen oxides ($NO_x$) and volatile organic compounds (VOCs) in the presence of solar radiation. Ozone formation in the troposphere is a rapid photochemical cycle and a non-linear process depending on the concentrations of precursors, meteorological parameters, and the sunlight intensity and spectral distribution [1, 2]. Briefly, it involves the photolysis of nitrogen dioxide ($NO_2$) by solar radiation to form nitric oxide (NO) and a ground-state oxygen atom (O).

$$NO_2 + h\nu \rightarrow NO + O \tag{1}$$

The major reaction forming $O_3$ in the troposphere includes the reaction of oxygen atom with oxygen molecule ($O_2$). Collision of recently formed $O_3$ with a third body (i.e., a molecule from the surrounding air) removes the excess energy of ozone and allows it to stabilize.

$$O + O_2 + M \rightarrow O_3 + M \tag{2}$$

$O_3$ is removed by the reaction with NO to reform $NO_2$;

$$O_3 + NO \rightarrow NO_2 + O_2 \tag{3}$$

However, reactions between NO and reactive radical species that are formed by the oxidation of reactive VOCs can also oxidize NO to $NO_2$ without the involvement of ozone,

$$NO \xrightarrow{HO_2\bullet,\ RO_2\bullet} NO_2 \tag{4}$$

Therefore, ozone concentration can increase.

Ozone is a strong oxidizing chemical. It has negative effects on human health, plants (e.g., decreased growth and biomass accumulation in plants, reduction in crop yields, damage to the leaves of plants, etc.), and materials. Textiles, fabrics, elastomers, paints, and other surface coatings can be damaged by ozone. There are number of studies (dosimetric, controlled human exposure, animal

**Correspondence:** Prof. F. Inal, Department of Chemical Engineering, Izmir Institute of Technology, Gulbahce-Urla, 35430 Izmir, Turkey
**E-mail:** fikretinal@iyte.edu.tr

**Abbreviations:** **ANN**, artificial neural network; **ARIMA**, autoregressive integrated moving average; **MAE**, mean absolute error; **MLP**, multi-layer perceptron; **MLR**, multiple linear regression; **MNLR**, multiple non-linear regression; **RMSE**, root mean squared error; **VOC**, volatile organic compound.

toxicologic, and epidemiological) in the literature on the health effects of ozone exposure [1, 3–5]. The health effects related to short-term ozone exposure include irritation of eyes and throat, adverse effects on pulmonary function, aggravation of respiratory symptoms, and increase in medication usage, hospital admissions and mortality. The long-term exposure can cause reduction in lung function development [6, 7]. Children, adults who are active outdoors [8], and people with preexisting respiratory illnesses are more sensitive to ozone than others. A recent study has shown that short-term inhalation of $PM_{2.5}$ and ozone at environmentally relevant concentrations causes acute conduit artery vasoconstriction in healthy adults [9].

Deterministic and stochastic methods are used in environmental modeling. Deterministic models require extensive data on reaction mechanisms, chemical kinetics, transport, and meteorological parameters. Artificial neural networks (ANNs) have recently been introduced as alternatives to conventional statistical methods for pollutant modeling due to the following advantages: ANNs make no prior assumptions concerning the data distribution; they can model highly non-linear functions; they can be retrained for better generalization whenever new or unseen data are available [10].

Yi and Prybutok [11] developed a neural network model for predicting daily maximum ozone levels using pollutant and meteorological data, and compared the neural network's performance with two statistical models: regression and Box–Jenkins autoregressive integrated moving average (ARIMA). The neural network model performed better than the regression and Box–Jenkins ARIMA models. Chattopadhyay and Chattopadhyay have applied autoregressive neural network (AR-NN) [12] and ARIMA [13] models to the monthly total ozone concentration over Kolkata, India. While the performance of AR-NN model was better compared with the ordinary autoregressive model of same order [12], ARIMA model in the form of ARIMA (0, 2, 2) had maximum prediction capacity among the three ARIMA and 11 AR-ANN models [13]. When only average daily meteorological data were used as input, Comrie [14] and Spellman [15] reported that the difference between the performances of neural network technique and regression model was not remarkable. In recent studies, ANN technique has been applied to predict tropospheric ozone levels in several European cities: five locations in United Kingdom (Central London, Harwell, Birmingham, Leeds, and Strath Vaich [15], and Bristol, Edinburgh, Eskdalemuir, Leeds, and Southampton [16]), Valencia, Spain [17, 18], Oporto, Portugal [19], and Orleans, France [20].

Istanbul is one of the world's largest cities with a population of about 12.6 million according to 2007 population census [21]. It also has the highest population density in Turkey with 2420 people/km$^2$. Air pollution is a major concern in Istanbul, since in addition to being the largest city in the country it contains 40% of the industrial facilities in Turkey [22]. The main industrial sectors are textile production, metal production, food processing, rubber, leather, chemical and petroleum products, machinery, and automotive. Recent meteorological evaluations, and tracer and trajectory studies have indicated that trans-boundary transport of air pollutants from Europe are also responsible for the poor air quality of Istanbul under specific weather conditions [23].

Different approaches have been used to model tropospheric ozone concentrations in Istanbul: non-linear time series method [24], regression model [25], fuzzy synthetic evaluation techniques [26], and cellular neural networks [27]. Ozcan et al. [27] have utilized genetically trained, multi-level cellular neural network to predict ozone values 24 h in advance. The input parameters for the model were meteorological and air pollutant data for the year 2003 (January to December). The correlation coefficients (*R*) between the predicted and measured ozone levels for the training and testing data sets were 0.62 and 0.57, respectively. In previous studies, little information is available on the effects of each input parameter on ozone concentrations in Istanbul.

Because of the adverse health effects of tropospheric ozone, it is indispensable to have an accurate model to forecast ozone concentrations. Our objective in this study was to predict next-day's maximum 1-h average ozone concentrations in Istanbul using multi-layer perceptron (MLP) ANN [10], the most frequently used ANN in atmospheric modeling, with larger datasets (i.e., air pollutant and meteorological data for the years 2003–2005). Additionally, sensitivity and pruning techniques were applied to find the effects of each input parameter and the simplified network architecture with these inputs, respectively. The performances of the MLPs were also compared with those of multiple linear regression (MLR) and multiple non-linear regression (MNLR) models.

## 2 Materials and methods

### 2.1 Site description and data

Istanbul (41.01°N, 28.58°E) is located in the northwest of Turkey (Fig. 1). Bosphorus strait divides the city into Asian and European parts. Due to its geographical location, northern and southern parts of Istanbul exhibit different meteorological characteristics [28]. The southern parts show general characteristics of the Mediterranean climate. However, in northern parts, Mediterranean type climate is modified by the cooler Black Sea and northerly colder air masses of maritime and continental origins. Therefore, the climate in this part of the city is described as having cooler temperatures in both winter and summer, and experiencing more rains compared to the south. The coldest months are January ($T_{\mathrm{ave}} = 6.1°\mathrm{C}$) and February ($T_{\mathrm{ave}} = 5.9°\mathrm{C}$) and the hottest months are July ($T_{\mathrm{ave}} = 23.8°\mathrm{C}$) and August ($T_{\mathrm{ave}} = 23.5°\mathrm{C}$). The average annual total precipitation is about 800 mm. The predominant winds are in the northeast direction in Istanbul.

The important air pollutant sources in Istanbul are residential heating, motor vehicle emissions, and industrial plants [29]. With respect to ozone precursors $NO_x$ and VOCs, the predominant sources are motor vehicle emissions and industrial plants for $NO_x$ emissions, and motor vehicle emissions and residential heating for the emissions of VOCs. As of June 2009, the number of motor vehicles registered to the traffic in Istanbul is about 2.8 million [30]. Natural gas, lignite, wood, and fuel-oil are the main fuels used for residential heating in winter season.

Air pollutant and meteorological data for three years (February to October for 2003 and 2004, and February to July for 2005) were used in the models developed in this study. Table 1 gives the input and output parameters. Air pollutant data were obtained from Kadikoy Air Quality Station operated by the Istanbul Metropolitan Municipality. The location of Kadikoy Air Quality station represents an urban site with traffic influence. Meteorological parameters were acquired from the nearest meteorological station, Goztepe Meteorology Station, operated by the State Meteorological Service. The distance between these two stations is about 6 km, and both of them are located in the Asian side.

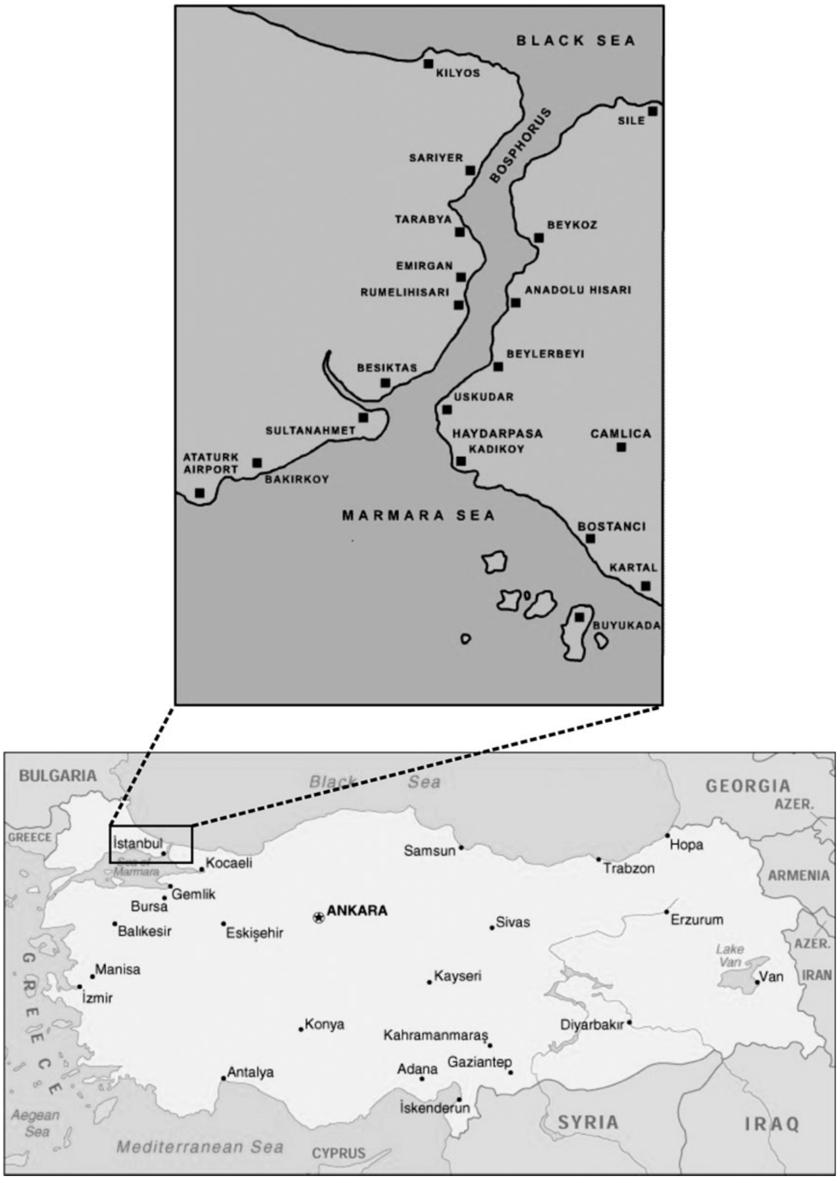

**Figure 1.** Map of Istanbul.

Due to the equipment maintenance some of the parameters were not available. If the values of any parameters were not measured more than 6 h for a given day, the entire row of data was removed from the dataset. Therefore, the total dataset used in this study was 578.

## 2.2 ANN model

ANNs are interconnected parallel systems. They consist of simple processing elements called neurons organized into layers [31, 32]. In contrast to traditional modeling approaches, ANNs are data driven and self-adaptive methods.

A feedforward, MLP type of ANNs was applied for the prediction of next-day's maximum 1-h average ozone concentrations. MLP consists of an input layer, one or more hidden layers, and an output layer (Fig. 2). Input quantities are fed into input layer neurons, and then distributed to all the neurons in the hidden layer without any computation. Each neuron in the hidden layer or output layer sums the weighted inputs received from the preceding layer to obtain its net input. Weights are adjustable parameters that determine the strength of the input signal. The output of a neuron in these layers (i.e., hidden or output) is computed by applying a transfer or activation function (usually non-linear) to its net input;

$$y_{\mathrm{j}} = f\left(\mathrm{net}_{\mathrm{j}}\right) = f\left(\sum_{j} w_{\mathrm{ij}} x_{\mathrm{i}} + b_{\mathrm{j}}\right) \tag{5}$$

where $y_{\mathrm{j}}$ is the output of the *j*th neuron, *f* the transfer function, $\mathrm{net}_{\mathrm{j}}$ the net input to the *j*th neuron, $w_{\mathrm{ij}}$ the connection weight from the *i*th neuron in the previous layer to the *j*th neuron in the current layer, $x_{\mathrm{i}}$ the input from the *i*th neuron to the *j*th neuron, and $b_{\mathrm{i}}$ is the bias [33].

Data normalization was performed before the training. Data were scaled to match the range of the hidden layer's transfer function. The ranges were 0–1 for the sigmoid transfer function and −1 to 1 for the hyperbolic tangent transfer function. The network output was denormalized to match the units of the desired response data.

**Table 1.** Input and output parameters used in the modeling study.

| Parameter | Timing | Unit |
|---|---|---|
| *Input* | | |
| $SO_2$ | Daily average[a] | $\mu g/m^3$ |
| $PM_{10}$ | Daily average[a] | $\mu g/m^3$ |
| CO | Daily average[a] | $\mu g/m^3$ |
| NO | Daily average[a] | $\mu g/m^3$ |
| $NO_2$ | Daily average[a] | $\mu g/m^3$ |
| $CH_4$ | Daily average[a] | $\mu g/m^3$ |
| Non-methane hydrocarbons (*n*MHCs) | Daily average[a] | $\mu g/m^3$ |
| $O_{3,ave}$ | Daily average[a] | $\mu g/m^3$ |
| $O_{3,max}$ | Daily maximum | $\mu g/m^3$ |
| Maximum temperature ($T_{max}$) | Daily maximum | °C |
| Average temperature ($T_{ave}$) | Daily average[a] | °C |
| Barometric pressure (BP) | Daily average[a] | mb |
| Relative humidity (RH) | Daily average[a] | % |
| Daily precipitation (DP) | Total daily | mm |
| Sunshine time (ST) | Total daily | h |
| Solar radiation (SR) | Daily average[a] | $cal/cm^2$ |
| Wind speed (WS) | Daily average[a] | m/s |
| Wind direction (WD) | Daily average[a] | N° |
| *Output* | | |
| Next day's maximum ozone concentration | Daily maximum | $\mu g/m^3$ |

[a] Average of the data measured at 07:00, 14:00, and 21:00 local standard time.

The learning of the network is achieved through the training process. Connection weights are updated using either supervised or unsupervised learning during the training. In supervised learning, input and output data (i.e., training set) are presented to the network. Backpropagation algorithm, the most commonly used supervised learning for MLPs, was applied to train the networks in this study. First step in backpropagation algorithm is to process the input in the forward direction to determine the output value of each neuron in the output layer. Network output for each neuron is then compared with the desired or target output, and the following error is calculated [33];

$$E = \sum_k \sum_n (d_{nk} - y_{nk})^2 \tag{6}$$

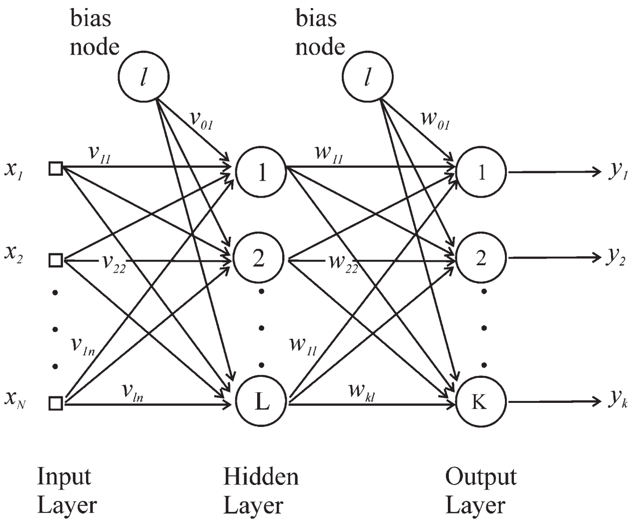

**Figure 2.** Feedforward, MLP type ANNs.

where $k$ is an index over the system output, $n$ is an index over the input patterns, $d$ is a component of the desired or target output vector **D**, and $y$ is a component of the network output vector **Y**. In the second step of backpropagation, this error is propagated in the backward direction from output layer to input layer to update the weights in each consecutive layer.

To speed up the learning and to avoid getting caught in local minima in the search for the optimal values of weights, we have used an adaptive search procedure Delta Bar Delta [33]. In this procedure, learning rates are adjusted continuously (i.e., learning rates are high when the learning curve (i.e., graph of output error versus iteration) is flat, and they are low when the learning curve oscillates) during training. If $\eta_{ij}$ is the learning rate for the weight $w_{ij}$, the update to each step size is

$$\Delta\eta_{ij}(n+1) = \begin{cases} k & \text{if} \quad S_{ij}(n-1)D_{ij}(n) > 0 \\ -b\eta_{ij}(n) & \text{if} \quad S_{ij}(n-1)D_{ij}(n) < 0 \\ 0 & \text{if} \quad \text{otherwise} \end{cases} \tag{7}$$

where $S_{ij}$ is the average of previous gradients and $D_{ij}$ is the current gradient. When the average of previous gradients and the current gradient have same sign, their product will be positive, which refers to slow convergence. Therefore, the step size increased arithmetically at each iteration by a constant. When the weight is oscillating (i.e., second case), the step size is decreased proportionally to its current value.Training can be performed either in batch or online modes. Weights are updated for each input sample in online training. However, in batch training, weight update occurs only after the presentation of the entire training set.

If the network is overtrained, it memorizes the training patterns, and thus poor generalization is obtained when the network is tested with unseen data. There are two stop criteria commonly used for ending the training process: stopping based on training-set error and stop criterion based on generalization (also known as early stopping or stopping with cross-validation) [33]. We have applied the stopping with cross-validation. For the best generalization of the network, training was stopped at the point of the smallest error in cross-validation set.

Once the neural network is trained, connection weights are fixed. The performance of the network is evaluated with the data (i.e., testing set) not used in training or cross-validation sets. The data from the years 2003–2004 were used in training (~80%), and 2005 were used in cross-validation (~5%) and testing (~15%). However, depending on the size of the available data set, different percentages of the total data can also be used in training, cross-validation, and testing.

The following network parameters were investigated and optimized during the development of the best network for the prediction of ozone levels in Istanbul: number of hidden layers (one or two hidden layers), number of neurons in each hidden layer (varied from 5 to 35 in each hidden layer), possible transfer functions (sigmoid or hyperbolic tangent), and training mode (batch or online).

## 3 Results and discussion

For the time period investigated in this study, daily maximum 1-h ozone concentrations occurred in the mid-afternoon (2–4 PM local standard time) while the lowest ozone levels were observed at about midnight or in the early morning in Istanbul. However, there were some cases in which ozone concentrations were high in the early morning hours. Similar behavior was also reported in previous studies [34]. Im et al. [34] have suggested that decreasing inversion heights in the early hours of the day leads to suppression of pollutants close to surface, and thus causing an increase in ozone concentrations. Summer months (June to August) had higher ozone levels. The daily maximum 1-h ozone concentrations were in the range of 18–163 $\mu g/m^3$ for June, 27–168 $\mu g/m^3$ for July, and 44–159 $\mu g/m^3$ for August. The information and alert thresholds for 1-h average ground-level ozone in National Ambient Air Quality Standards of Turkey are 180 and 240 $\mu g/m^3$, respectively [35]. The target value for the daily maximum 8-h average ozone concentration to be achieved by 2022 is 120 $\mu g/m^3$ (not to be exceeded more than 25 days per calendar year; averaged over 3 years).

The previous studies have shown that the performance of ANNs could be improved by including persistence information as an input [14, 20, 36]. Therefore, current day's ozone concentrations (daily average and maximum 1-h concentrations) were used as input parameters. Initially, the total number of inputs was eighteen (i.e., nine pollutant parameters and nine meteorological parameters).

The performances of the models were evaluated with the following statistical indicators: Pearson product moment correlation coefficient (*R*), mean absolute error (MAE), root mean squared error (RMSE), and index of agreement or Willmott's Index ($d_2$):

$$R = \frac{\dfrac{\sum_i (d_i - \bar{d})(y_i - \bar{y})}{N}}{\sqrt{\dfrac{\sum_i (d_i - \bar{d})^2}{N}}\sqrt{\dfrac{\sum_i (y_i - \bar{y})^2}{N}}} \tag{8}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_i |d_i - y_i| \tag{9}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_i (d_i - y_i)^2} \tag{10}$$

$$d_2 = 1 - \frac{\sum_i |y_i - d_i|^2}{\sum_i \left(|y_i - \bar{d}| + |d_i - \bar{d}|\right)^2} \tag{11}$$

Hundreds of networks were tested to obtain the best prediction performance. The optimum MLP architecture was found to be 1-hidden layer with 24 neurons (**18-24-1**). The transfer function and training mode were hyperbolic tangent and batch, respectively.

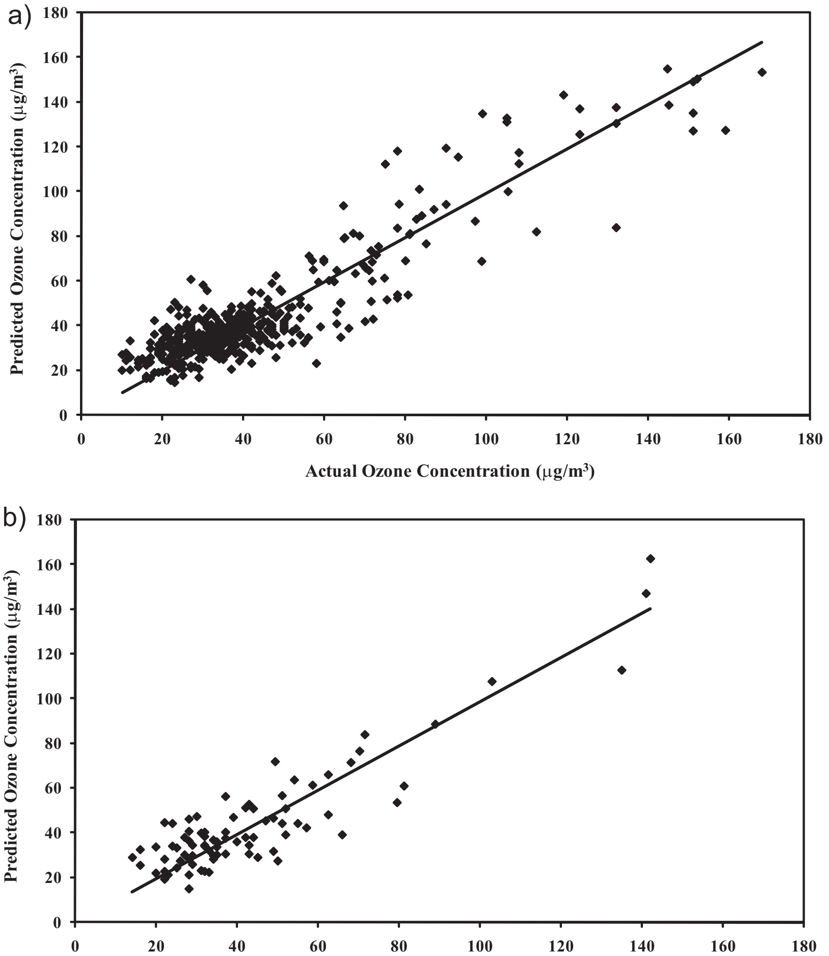

**Figure 3.** Scatter plots for the MLP model with 18 inputs (a) training (b) testing.

**Table 2.** Performance summary of the MLP model with 18 inputs.

| Performance indicator | Training | Testing |
|---|---|---|
| MAE ($\mu$g/m$^3$) | 8.14 | 8.78 |
| RMSE ($\mu$g/m$^3$) | 11.07 | 11.15 |
| $R$ | 0.90 | 0.90 |
| $d_2$ | 0.95 | 0.95 |

The performance of the network did not improve considerably with increases in number of hidden layer and number of neuron in each hidden layer. Since the larger networks require more free parameters to solve a given problem and may overfit the data, a network with low complexity was preferred.

Scatter plots of daily maximum actual ozone concentrations against actual concentrations are given in Fig. 3 for the training and testing datasets. The actual and predicted concentrations were in good agreement. The deviations from the diagonal were random and not systematic. The correlation coefficient, MAE, RMSE, and $d_2$ for the training data were 0.90, 8.14 $\mu$g/m$^3$, 11.07 $\mu$g/m$^3$, and 0.95, respectively (Tab. 2). When MLP model was tested with unseen data, the correlation coefficient and $d_2$ were 0.90 and 0.95, respectively, which indicates that the model did not over-train or memorize the data patterns. MAE and RMSE for the testing data were 8.78 and 11.15 $\mu$g/m$^3$, respectively (Tab. 2). A good generalization performance was obtained since there were not significant differences between

**Table 3.** Performances of the MLP models having different time lags of $O_{3,max}$ as inputs.

| Inputs | Optimum MLP architecture[a] | MAE ($\mu$g/m$^3$) | RMSE ($\mu$g/m$^3$) | $R$ | $d_2$ |
|---|---|---|---|---|---|
| 18 Inputs (including $O_{3,max}(t-1)$) | 18-24-1 | 8.78 | 11.15 | 0.90 | 0.95 |
| 19 Inputs (including $O_{3,max}(t-1)$ and $O_{3,max}(t-2)$) | 19-26-1 | 9.38 | 11.53 | 0.89 | 0.94 |
| 20 Inputs (including $O_{3,max}(t-1)$, $O_{3,max}(t-2)$, and $O_{3,max}(t-3)$) | 20-25-1 | 11.52 | 13.97 | 0.90 | 0.91 |

[a] Transfer function: hyperbolic tangent.
Training mode: batch.

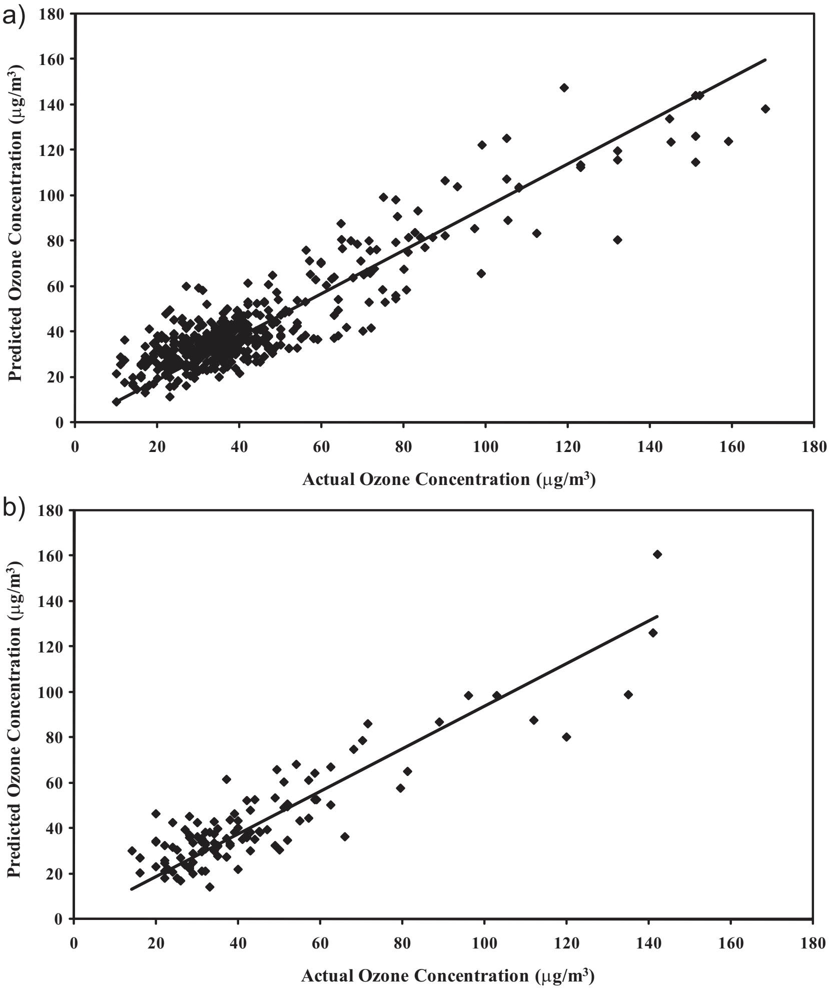

**Figure 4.** Scatter plots for the MLR model with 18 inputs (a) training (b) testing.

**Table 4.** Performance summary of the MLR model with 18 inputs.

| Performance indicator | Training | Testing |
|---|---|---|
| MAE ($\mu g/m^3$) | 8.12 | 8.65 |
| RMSE ($\mu g/m^3$) | 10.89 | 11.35 |
| $R$ | 0.90 | 0.89 |
| $d_2$ | 0.95 | 0.94 |

**Table 5.** Performance summary of the MNLR model with 18 inputs.

| Performance indicator | Training | Testing |
|---|---|---|
| MAE ($\mu g/m^3$) | 8.11 | 8.09 |
| RMSE ($\mu g/m^3$) | 11.94 | 11.86 |
| $R$ | 0.89 | 0.88 |
| $d_2$ | 0.94 | 0.92 |

the values of performance indicators for the training and testing sets.

Ozcan et al. [27] have reported the values of MAE and RMSE as 6.32 and 8.70 $\mu g/m^3$, respectively for the prediction of next-day daily mean ozone concentrations in Istanbul using cellular neural network approach. However, their correlation coefficient ($R = 0.57$) was lower than the one we obtained in this study. Dutot et al. [20] have utilized MLP network combined to a neural classifier for the forecasting of next-day maximum hourly-mean ozone concentrations. The network inputs were the model output of the weather predictions and persistence variables. They obtained higher MAE and RMSE ($MAE = RMSE = 15\ \mu g/m^3$). The index of agreement was 0.92. In a similar study, Salazar-Ruiz et al. [37] have used meteorological data, precursor concentrations, and persistence information as inputs to predict next-day maximum tropospheric ozone levels. $R$, $d_2$, and RMSE for the MLP model developed were 0.74, 0.85, and 9.43 ppb ($1\ ppb = 1.96\ \mu g/m^3$ at 25°C), respectively.

Although persistence information in the form of 1-day time lag was used as an input in the MLP model, we have calculated the autocorrelation coefficients for the daily maximum ozone concentration ($O_{3,\ max}$) up to time lag of 3 days using:

$$r = \frac{\sum (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum (x_t - \bar{x})^2} \tag{12}$$

where $x_t$ and $x_{t-k}$ are the paired values, and $k$ is the lag. The autocorrelation coefficients for $O_{3,max}$ were 0.73, 0.62, and 0.51 for the lag of 1, 2, and 3 days, respectively. Therefore, we have decided to check

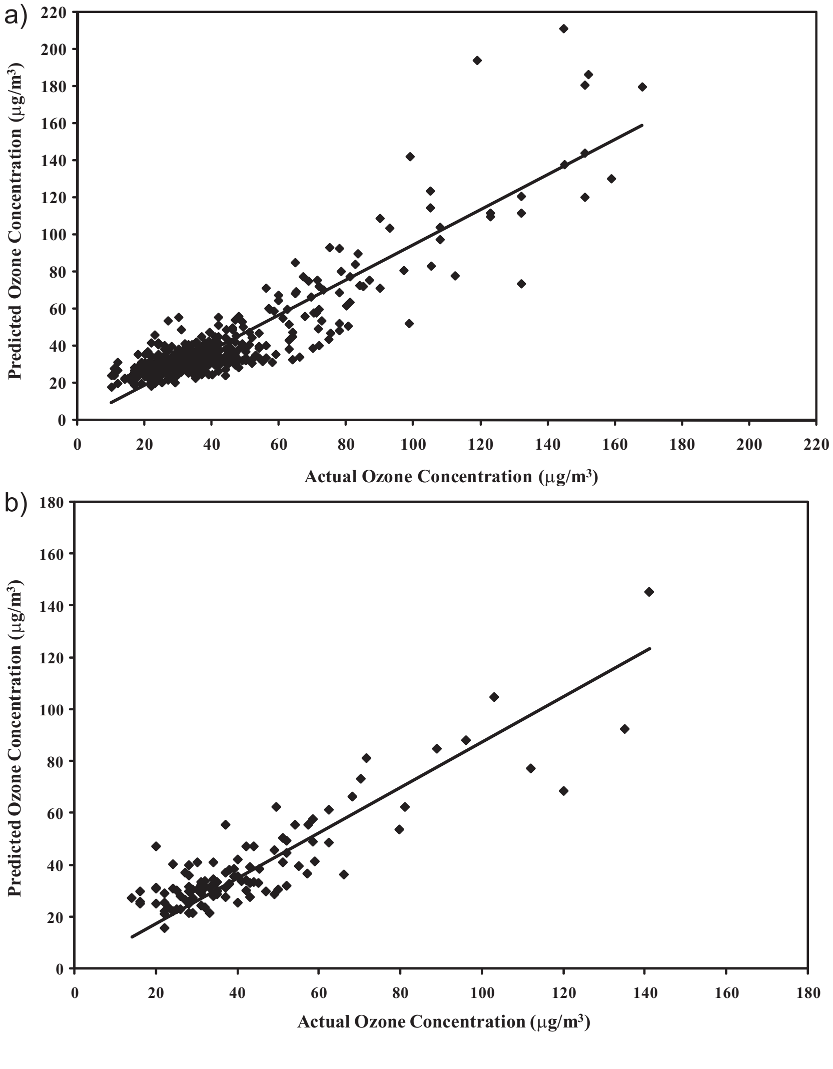

**Figure 5.** Scatter plots for the MNLR model with 18 inputs (a) training (b) testing.

**Table 6.** Sensitivity factors for the input parameters.

| Input | Sensitivity factor |
|---|---|
| BP | 0.12 |
| $CH_4$ | 0.13 |
| CO | 0.14 |
| WS | 0.15 |
| *n*MHCs | 0.25 |
| $NO_2$ | 0.29 |
| $SO_2$ | 0.32 |
| RH | 0.50 |
| DP | 0.60 |
| SR | 0.86 |
| WD | 0.94 |
| ST | 1.03 |
| $T_{max}$ | 1.08 |
| $PM_{10}$ | 1.37 |
| $T_{ave}$ | 1.43 |
| NO | 1.48 |
| $O_{3,ave}$ | 4.02 |
| $O_{3,max}$ | 5.78 |

the performances of MLP models containing additional 2- and 3-day lags as input. Results have indicated that there were not any improvements in the performances of the MLP models when this additional input was used (Tab. 3). Consequently, only 1-day lag information was utilized in this study.

We compared the performance of the MLP model with MLR and MNLR models. In regression modeling, a new test dataset was formed by the addition of cross-validation data to the test data.

The coefficients of the MLR model were obtained by the ordinary least squares procedure (SPSS v.7.0) without stepwise regression. The model equation was:

$$\begin{aligned}[O_3] = &-109.941 + 0.1166[\mathrm{BP}] + 0.00064[\mathrm{CH_4}] - 0.00086[\mathrm{CO}]\\ &+ 0.068[\mathrm{DP}] + 0.0049[n\mathrm{MHC}] + 0.0026[\mathrm{NO}]\ - 0.0207[\mathrm{NO_2}]\\ &+ 0.724[O_{3,}\,\mathrm{ave}] + 0.438[O_{3,\mathrm{max}}] + 0.046[PM_{10}] + 0.007[\mathrm{RH}]\\ &- 0.0069[\mathrm{SR}] + 0.0258[\mathrm{SO_2}]\ - 0.2203[\mathrm{ST}] + 0.545[T_{\mathrm{ave}}]\\ &- 0.2366[T_{\mathrm{max}}] - 0.01098[\mathrm{WD}] - 1.474[\mathrm{WS}]\end{aligned} \quad (13)$$

Scatter plots of actual ozone concentrations against actual concentrations are given in Fig. 4 for the training and testing data sets, respectively. The deviations from the diagonal were not systematic. The performance of the MLR model was also good for predicting next-day maximum ozone concentrations. Table 4 summarizes the performance indicators for the MLR model. *R*, MAE, RMSE, and $d_2$ for the testing data were 0.89, 8.65 μg/$m^3$, 11.35 μg/$m^3$, and 0.94, respectively. Tecer et al. [25] have also investigated ozone forecasting in Istanbul using MLR model, and obtained an $R^2$ (coefficient of determination) [38] value of 0.715.

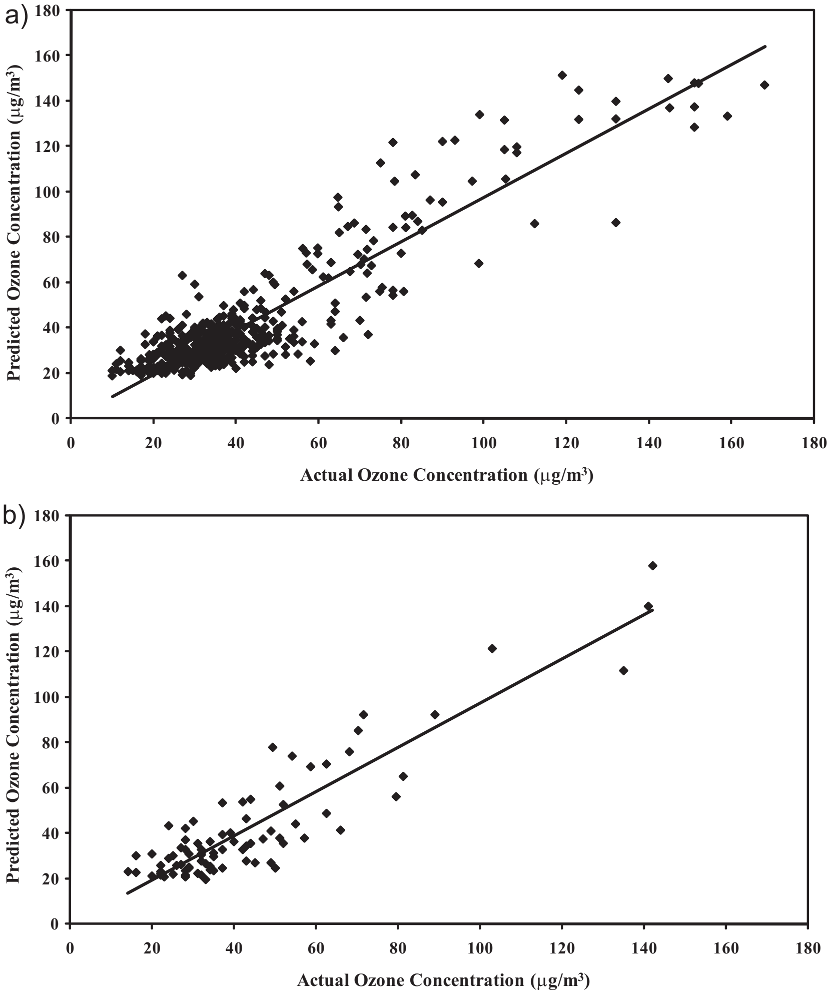

**Figure 6.** Scatter plots for the MLP model with nine inputs (a) training (b) testing.

**Table 7.** Performance summary of the MLP model with nine inputs.

| Performance indicator | Training | Testing |
|---|---|---|
| MAE ($\mu g/m^3$) | 8.45 | 9.34 |
| RMSE ($\mu g/m^3$) | 11.46 | 11.64 |
| $R$ | 0.90 | 0.90 |
| $d_2$ | 0.95 | 0.95 |

A modified form of MNLR equation of Chattopadhyay and Chattopadhyay [39] was also used to predict ozone concentrations:

$$\ln(y) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots \quad (14)$$

The regression parameters were estimated using the Levenberg–Marquardt method. The model equation was:

$$\begin{aligned}\ln[O_3] = {} & 0.178 + 0.0027[\mathrm{BP}] + 0.000085[\mathrm{CH_4}] - 0.000064[\mathrm{CO}] \\ & + 0.0036[\mathrm{DP}] + 0.000081[n\mathrm{MHC}] - 0.00049\ [\mathrm{NO}] - 0.0013[\mathrm{NO_2}] \\ & + 0.0091[\mathrm{O_{3,ave}}] + 0.008[\mathrm{O_{3,max}}] + 0.0018[PM_{10}] - 0.00046[\mathrm{RH}] \\ & + 0.0032[\mathrm{SR}] - 0.0012\ [\mathrm{SO_2}] - 0.0021[\mathrm{ST}] + 0.0045[T_{\mathrm{ave}}] \\ & + 0.0031[T_{\max}] - 0.00026[\mathrm{WD}] - 0.0037[\mathrm{WS}]\end{aligned} \quad (15)$$

For the MNLR model, scatter plots of actual ozone concentrations against actual concentrations are given in Fig. 5 for the training and testing datasets. $R$, MAE, RMSE, and $d_2$ for the testing data were 0.88, 8.09 $\mu g/m^3$, 11.86 $\mu g/m^3$, and 0.92, respectively (Tab. 5). The values of $R$ and $d_2$ were slightly lower compared to those obtained with MLP and MLR models. Elkamel et al. [40] have reported the average, maximum, and minimum errors for the testing set as 20.04, 188.6, and 0.229%, respectively for the forecasting of ozone levels in Kuwait using a similar MNLR model.

In order to understand or interpret the results obtained from MLP models, sensitivity and pruning analyses are usually carried out. There are different methods used for sensitivity analysis [17]. In the delta error method, the changes in training error that would be obtained if an input were removed from the model are evaluated. However, in the average absolute gradient method, an input is perturbed, and then the model outputs are monitored. In this study, to find the effects of each input parameter on next day's maximum ozone concentration, the network was first trained and the connection weights were fixed. After that, one by one, each input parameter was randomly perturbed around its mean value while the other inputs were kept at their mean values, and then the change in the output was measured. The input perturbation was done by adding a random value of a known variance to each

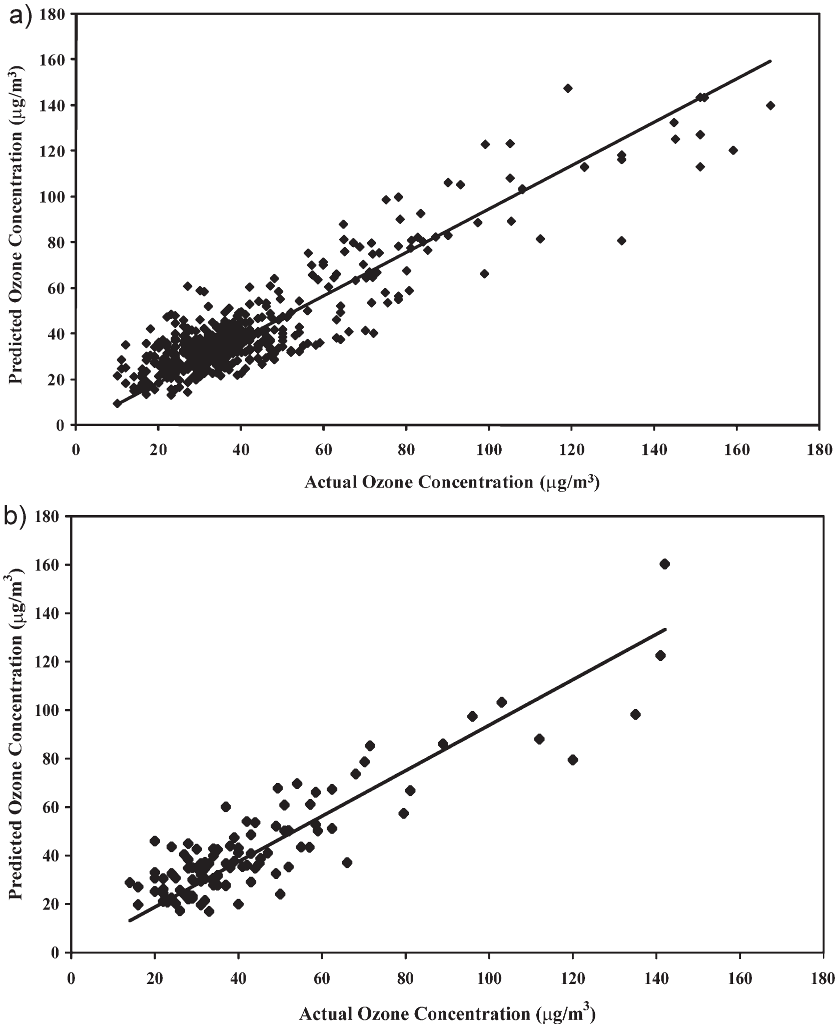

**Figure 7.** Scatter plots for the MLR model with nine inputs (a) training (b) testing.

sample and computing the output. The sensitivity factor for input *k* is given as:

$$S_k = \frac{\sum_{p=1}^{P}\sum_{i=1}^{o}\left(y_{ip} - \overline{y_{ip}}\right)^2}{\sigma_k^2} \tag{16}$$

where $\overline{y_{ip}}$ is the *i*th output obtained with the fixed weights for the *p*th pattern, *o* the number of network outputs, *P* the number of patterns, and $\sigma_k^2$ is the variance of the input perturbation [33].

The sensitivity factors are given in Tab. 6. The most important input parameters were found to be persistence information (current day's maximum and average ozone concentrations), NO concentration, average temperature, $PM_{10}$, maximum temperature, sunshine time, wind direction, and solar radiation. The previous studies have also indicated the significance of persistence information in ozone prediction [20, 36]. $NO_x$ (NO and $NO_2$) species are the ozone precursors. Since the formation of ozone is a photochemical reaction, reaction rate is influenced by the temperature and the intensity of solar radiation. The higher sensitivity factor obtained for the wind direction indicates the importance of transport phenomena for ozone levels in Istanbul. $PM_{10}$ concentrations affect the solar radiation intensity. The sensitivity factors for hydrocarbon species (i.e., *n*MHCs and $CH_4$) were relatively low.

The input parameters with smaller sensitivity factors (i.e., DP, RH, $SO_2$, $NO_2$, *n*MHCs, WS, CO, $CH_4$, and BP) were removed in Pruning tests, and the simplified networks with fewer connection weights were evaluated again for their prediction performances. When only nine input parameters were used for the forecasting of daily maximum ozone concentrations, the optimum MLP architecture was found to be 1-hidden layer with 24 neurons (**9-24-1**). The transfer function was hyperbolic tangent. The better prediction performances were obtained with batch training mode.

Scatter plots of actual and predicted maximum ozone concentrations are shown in Fig. 6 for the training and testing sets. The predictions of the simplified network were also consistent with both training and testing data. *R*, MAE, RMSE, and $d_2$ were 0.90, 8.45 $\mu g/m^3$, 11.46 $\mu g/m^3$, and 0.95 for the training data, and 0.90, 9.34 $\mu g/m^3$, 11.64 $\mu g/m^3$, and 0.95 for the testing data, respectively (Tab. 7). The removal of input parameters with lower sensitivity factors was supported by not having significant differences in performance indicators for 18- and 9-input networks.

We have also checked the performance of MLR and MNLR models with nine inputs. Scatter plots of actual and predicted daily maximum ozone concentrations for the training and testing sets are shown in Fig. 7 for the MLR model and in Fig. 8 for the MNLR model. In general, MLR and MNLR models with nine inputs performed as good as the models with 18 inputs. *R*, MAE, RMSE, and $d_2$ for the

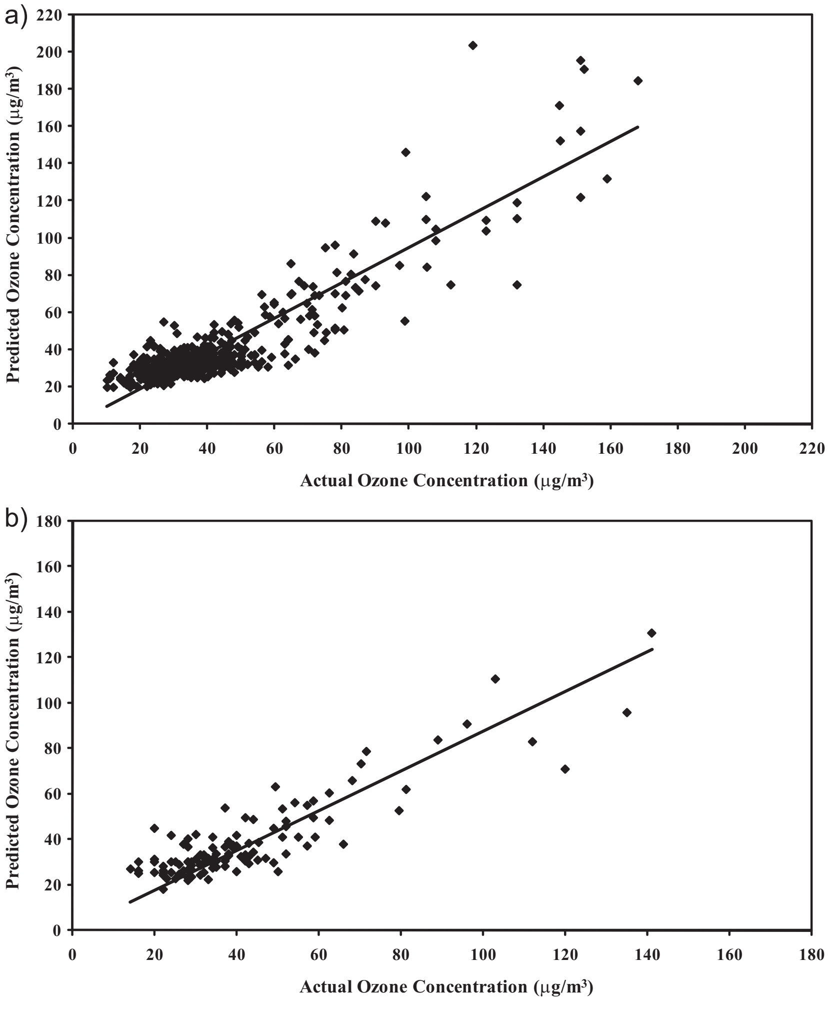

**Figure 8.** Scatter plots for the MNLR model with nine inputs (a) training (b) testing.

**Table 8.** Performance summary of the MLR model with nine inputs.

| Performance indicator | Training | Testing |
|---|---|---|
| MAE ($\mu g/m^3$) | 8.18 | 8.66 |
| RMSE ($\mu g/m^3$) | 10.98 | 11.49 |
| $R$ | 0.90 | 0.89 |
| $d_2$ | 0.95 | 0.94 |

**Table 9.** Performance summary of the MNLR model with nine inputs.

| Performance indicator | Training | Testing |
|---|---|---|
| MAE ($\mu g/m^3$) | 8.18 | 8.15 |
| RMSE ($\mu g/m^3$) | 11.93 | 11.52 |
| $R$ | 0.89 | 0.88 |
| $d_2$ | 0.94 | 0.91 |

testing set were 0.89, 8.66 $\mu g/m^3$, 11.49 $\mu g/m^3$, and 0.94 for the MLR model (Tab. 8), and 0.88, 8.15 $\mu g/m^3$, 11.52 $\mu g/m^3$, and 0.91 for the MNLR model (Tab. 9), respectively.

## 4 Conclusions

MLP ANN and regression (i.e., MLR and MNLR) modeling approaches were successfully applied to predict next day's maximum 1-h average ozone concentrations in Istanbul. The most significant input parameters were found by the sensitivity analysis. A simplified MLP model was obtained by removing the input parameters with lower sensitivity factors in Pruning tests. It has been shown that only nine inputs (current day's maximum and average ozone concentrations, NO concentrations, average temperature, $PM_{10}$, maximum temperature, sunshine time, wind direction, and solar radiation) were enough to predict next day's maximum ozone concentrations without any significant performance loss. There were also good agreements between the actual data and regression model results with nine inputs. Although slightly lower values of $R$ and $d_2$ were obtained for MNLR models, the performances of MLP and regression models with both 18 and 9 inputs were comparable for predicting the ozone concentrations in Istanbul based on the four statistical indicators considered.

### Acknowledgments

The financial support from Izmir Institute of Technology is gratefully acknowledged. We would like to thank M. Dogan (Istanbul Metropolitan Municipality) for providing air pollutant data, and anonymous reviewers for their valuable comments and suggestions.

*The author has declared no conflict of interest.*

## References

[1] US EPA (United States Environmental Protection Agency), *Air Quality Criteria for Ozone and Related Photochemical Oxidations*, EPA 600/R-05/004aF, Washington, DC **2006**.

[2] K. Y. Kondratyev, C. A. Varotsos, Global Tropospheric Ozone Dynamics I, *Environ. Sci. Pollut. Res.* **2001**, *8*, 57–62.

[3] M. Lippmann, Health Effects of Ozone: A Critical Review, *J. Air Waste Manage. Assoc.* **1989**, *39*, 672–695.

[4] B. Brunekreef, S. T. Holgate, Air Pollution and Health, *Lancet* **2002**, *360*, 1233–1242.

[5] J. A. Bernstein, N. Alexis, C. Barnes, I. L. Bernstein, J. A. Bernstein, A. Nel, D. Peden, et al., Health Effects of Air Pollution, *J. Allergy Clin. Immunol.* **2004**, *114*, 1116–1123.

[6] WHO Europe (World Health Organization Europe), *Health Aspects of Air Pollution*, Brussels **2004**.

[7] I. B. Tager, J. Balmes, F. Lurmann, L. Ngo, S. Alcorn, N. Kunzli, Chronic Exposure to Ambient Ozone and Lung Function in Young Adults, *Epidemiology* **2005**, *16*, 751–759.

[8] M. Brauer, J. R. Brook, Ozone Personal Exposures and Health Effects for Selected Groups Residing in the Fraser Valley, *Atmos. Environ.* **1997**, *31*, 2113–2121.

[9] R. D. Brook, J. R. Brook, B. Urch, R. Vincent, S. Rajagopalan, F. Silverman, Inhalation of Fine Particulate Air Pollution and Ozone Causes Acute Arterial Vasoconstriction in Healthy Adults, *Circulation* **2002**, *105*, 1534–1536.

[10] M. W. Gardner, S. R. Dorling, Artificial Neural Networks (The Multilayer Perceptron) – A Review of Applications in the Atmospheric Sciences, *Atmos. Environ.* **1998**, *32*, 2627–2636.

[11] J. Yi, V. R. Prybutok, A Neural Network Model Forecasting for Prediction of Daily Maximum Ozone Concentration in an Industrialized Urban Area, *Environ. Pollut.* **1996**, *92*, 349–357.

[12] G. Chattopadhyay, S. Chattopadhyay, Autoregressive Forecast of Monthly Total Ozone Concentration: A Neurocomputing Approach, *Comput. Geosci.* **2009**, *35*, 1925–1932.

[13] G. Chattopadhyay, S. Chattopadhyay, Univariate Approach to the Monthly Total Ozone Time Series Over Kolkata, India: Autoregressive Integrated Moving Average (ARIMA) and Autoregressive Neural Network (AR-NN) Models, *Int. J. Remote Sens.* **2010**, *31*, 575–583.

[14] A. C. Comrie, Comparing Neural Networks and Regression Models for Ozone Forecasting, *J. Air Waste Manage. Assoc.* **1997**, *47*, 653–663.

[15] G. Spellman, An Application of Artificial Neural Networks to the Prediction of Surface Ozone Concentrations in the United Kingdom, *Appl. Geogr.* **1999**, *19*, 123–136.

[16] M. W. Gardner, S. R. Dorling, Statistical Surface Ozone Models: An Improved Methodology to Account for Non-linear Behavior, *Atmos. Environ.* **2000**, *34*, 21–34.

[17] O. Pastor-Barcenas, E. Soria-Olivas, J. D. Martin-Guerrero, G. Camps-Valls, J. L. Carrasco-Rodriguez, S. D. Valle-Tascon, Unbiased Sensitivity Analysis and Pruning Techniques in Neural Networks for Surface Ozone Modeling, *Ecol. Modell.* **2005**, *182*, 149–158.

[18] J. Gomez-Sanchis, J. D. Martin-Guerrero, E. Soria-Olivas, J. Vila-Frances, J. L. Carrasco, S. D. Valle-Tascon, Neural Networks for Analyzing the Relevance of Input Variables in the Prediction of Tropospheric Ozone Concentration, *Atmos. Environ.* **2006**, *40*, 6173–6180.

[19] S. I. V. Sousa, F. G. Martins, M. C. M. Alvim-Ferraz, M. C. Pereira, Multiple Linear Regression and Artificial Neural Networks Based on Principal Components to Predict Ozone Concentrations, *Environ. Modell. Softw.* **2007**, *22*, 97–103.

[20] A.-L. Dutot, J. Rynkiewicz, F. E. Steiner, J. Rude, A 24-h Forecast of Ozone Peaks and Exceedance Levels Using Neural Classifiers and Weather Predictions, *Environ. Modell. Softw.* **2007**, *22*, 1261–1269.

[21] TURKSTAT (Turkish Statistical Institute), *Address Based Population Registration System 2007 Population Census Result*, Press Release, Number 9, Ankara **2008**.

[22] E. Durukal, M. Erdik, E. Uckan, Earthquake Risk to Industry in Istanbul and Its Management, *Nat. Hazards* **2008**, *44*, 199–212.

[23] T. Kindap, Identifying the trans-Boundary Transport of Air Pollutants to the City of Istanbul under Specific Weather Conditions, *Water Air Soil Pollut.* **2008**, *189*, 279–289.

[24] K. Kocak, L. Saylan, O. Sen, Nonlinear Time Series Prediction of $O_3$ Concentration in Istanbul, *Atmos. Environ.* **2000**, *34*, 1267–1271.

[25] L. H. Tecer, F. Erturk, O. Cerit, Development of a Regression Model to Forecast Ozone Concentration in Istanbul City, Turkey, *Fresenius Environ. Bull.* **2003**, *12*, 1133–1143.

[26] G. Onkal-Engin, I. Demir, H. Hiz, Assessment of Urban Air Quality in Istanbul Using Fuzzy Synthetic Evaluation, *Atmos. Environ.* **2004**, *38*, 3809–3815.

[27] H. K. Ozcan, E. Bilgili, U. Sahin, O. N. Ucan, C. Bayat, Modeling of Tropospheric Ozone Concentrations Using Genetically Trained Multi-level Cellular Neural Networks, *Adv. Atmos. Sci.* **2007**, *24*, 907–914.

[28] Y. Ezber, O. L. Sen, T. Kindap, M. Karaca, Climatic Effects of Urbanization in Istanbul: A Statistical and Modeling Analysis, *Int. J. Climatol.* **2007**, *27*, 667–679.

[29] IMM (Istanbul Metropolitan Municipality), *Istanbul Air Quality Strategy*, Istanbul Metropolitan Municipality Environmental Protection and Control Department, Istanbul **2009**.

[30] TURKSTAT (Turkish Statistical Institute), *Statistics for Motor Vehicles*, Press Release, Number 149, Ankara **2009**.

[31] M. M. Nelson, W. T. Illingworth, *A Practical Guide to Neural Nets*, Addison-Wesley, Reading, Massachusetts **1994**.

[32] L. Fausett, *Fundamentals of Neural Networks Architectures, Algorithms, and Applications*, Prentice-Hall, New Jersey **1994**.

[33] J. C. Principe, N. R. Euliano, W. C. Lefebvre, *Neural and Adaptive Systems: Fundamentals Through Simulations*, John Wiley & Sons, New York **1999**.

[34] U. Im, M. Tayanc, O. Yenigun, Analysis of Major Photochemical Pollutants with Meteorological Factors for High Ozone Days in Istanbul, Turkey, *Water Air Soil Pollut.* **2006**, *175*, 335–359.

[35] Official Journal (Official Journal of the Turkey), *Air Quality Assessment and Management Regulation*, Ministry of Environment and Forestry, Number 26898, Ankara **2008**.

[36] M. Cai, Y. Yin, M. Xie, Prediction of Hourly Air Pollutant Concentrations Near Urban Arterials Using Artificial Neural Network Approach, *Transport. Res. Part D* **2009**, *14*, 32–41.

[37] E. Salazar-Ruiz, J. B. Ordieres, E. P. Vergara, S. F. Capuz-Rizo, Development and Comparative Analysis of Tropospheric Ozone Prediction Models Using Linear and Artificial Intelligence-based Models in Mexicali, Baja California (Mexico) and Calexico, California (US), *Environ. Modell. Softw.* **2008**, *23*, 1056–1069.

[38] D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, California **1995**.

[39] S. Chattopadhyay, G. Chattopadhyay, Identification of the Best Hidden Layer Size for Three-layered Neural Net in Predicting Monsoon Rainfall in India, *J. Hydroinf.* **2008**, *10*, 181–188.

[40] A. Elkamel, S. Abdul-Wahab, W. Bouhamra, E. Alper, Measurement and Prediction of Ozone Levels Around a Heavily Industrialized Area: A Neural Network Approach, *Adv. Environ. Res.* **2001**, *5*, 47–59.